RSAC 2025 Conference

Many Voices.
One Community.

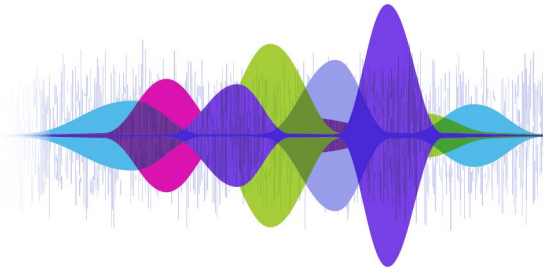
SESSION ID: USENIX-M02B

Adversarial Illusions in Multi-Modal Embeddings

Tingwei Zhang

PhD Student
Cornell Tech
https://ztingwei.com/

Disclaimer



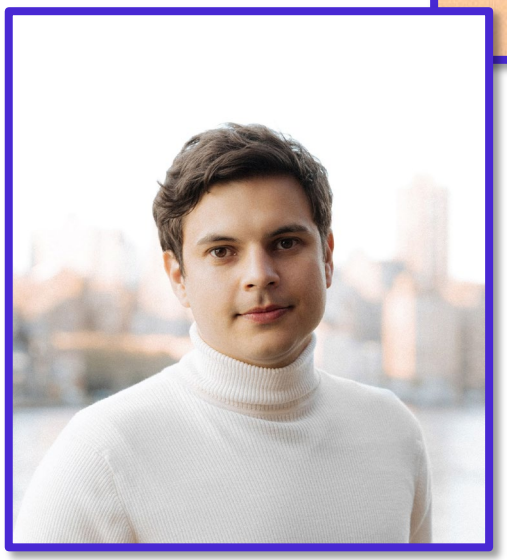
Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Vitaly Shmatikov

Eugene Bagdasaryan



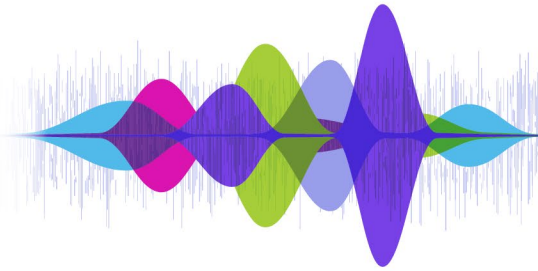
Rishi Jha

Fnu Suya

Collin Zhang

RSAC 2025 Conference

Multi-modal Models Are All the Rage



Meta ImageBind: An AI Model That Mimics Human Perception

Amazon Titan Embeddings for enhanced content recommendations to power 1:1 personalization

Create Your Own Multimodal Search Engine Using Google's Vertex AI

Meta Releases Llama 3.2 with Vision, Voice, and Open Customizable Models

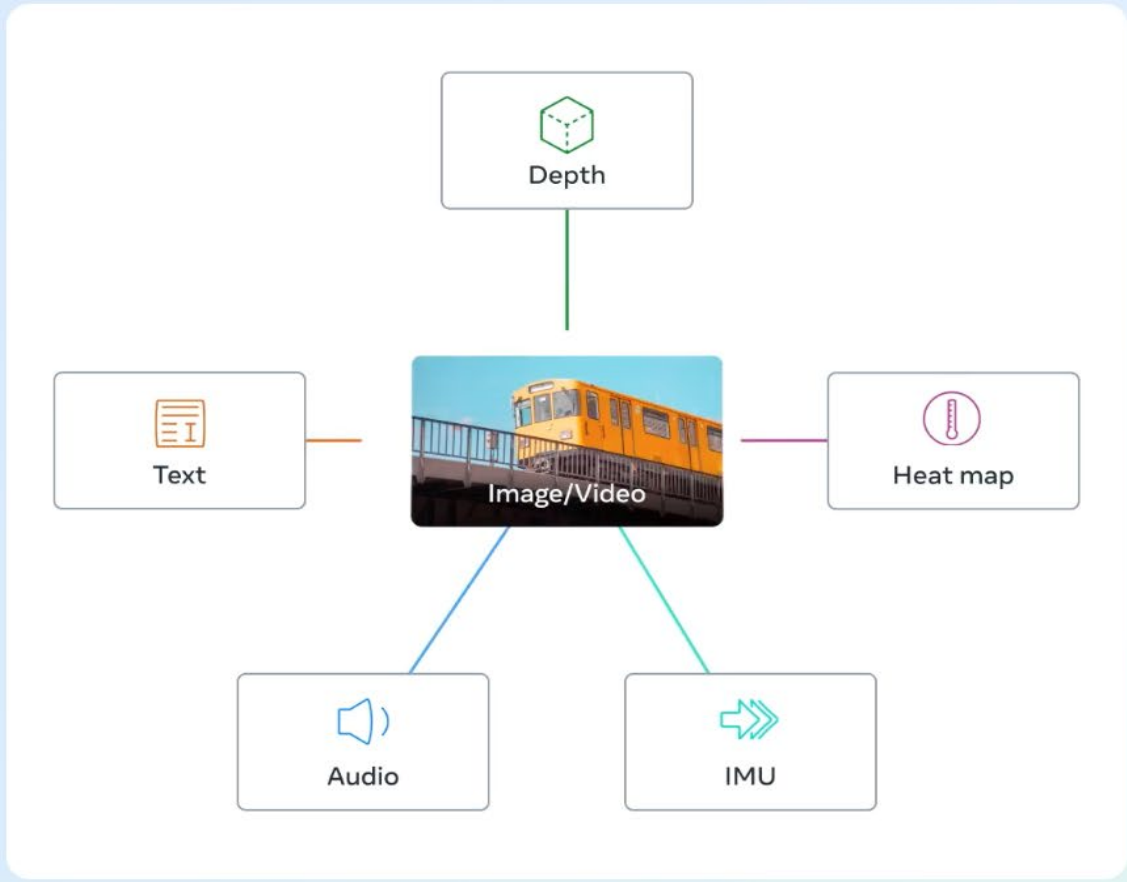
RSAC 2025 Conference

Multi-modal Models Are All the Rage

ImageBind: a new way to ‘link’ AI across the senses

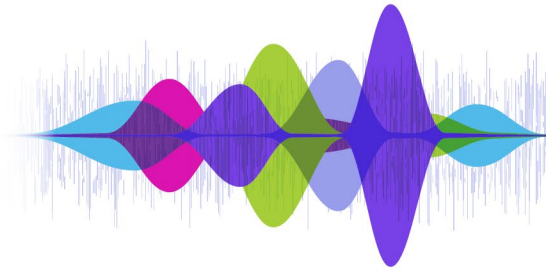
Introducing ImageBind, the first AI model capable of binding data from six modalities at once, without the need for explicit supervision. By recognizing the relationships between these modalities — images and video, audio, text, depth, thermal and inertial measurement units (IMUs) — this breakthrough helps advance AI by enabling machines to better analyze many different forms of information, together.

Explore the demo to see ImageBind's capabilities across image, audio and text modalities.



RSAC 2025 Conference

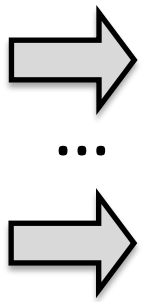
How (Some) Multi-Modal Models Work



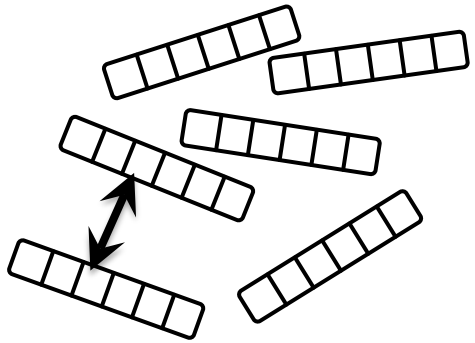
If these are compromised…

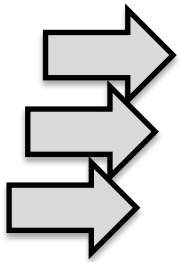
Text, images, sounds…

Encoders

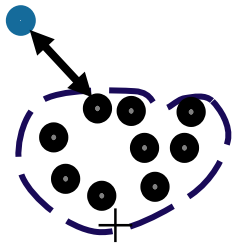


Vector embeddings

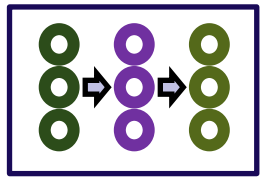




Retrieval, classification, generation



Diffusion



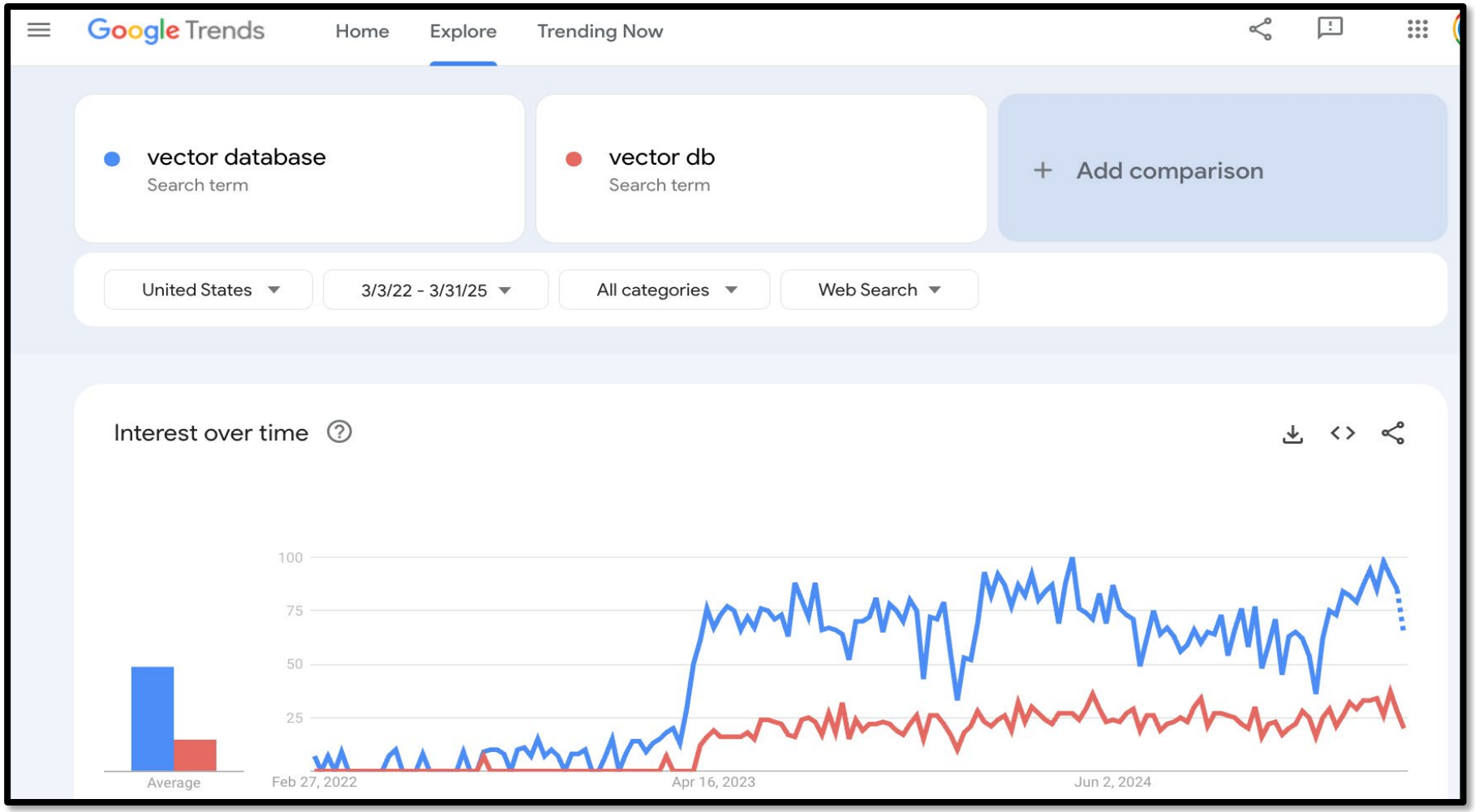
Language models

alignment:
related inputs → similar vectors

… all downstream tasks are compromised

RSAC 2025 Conference

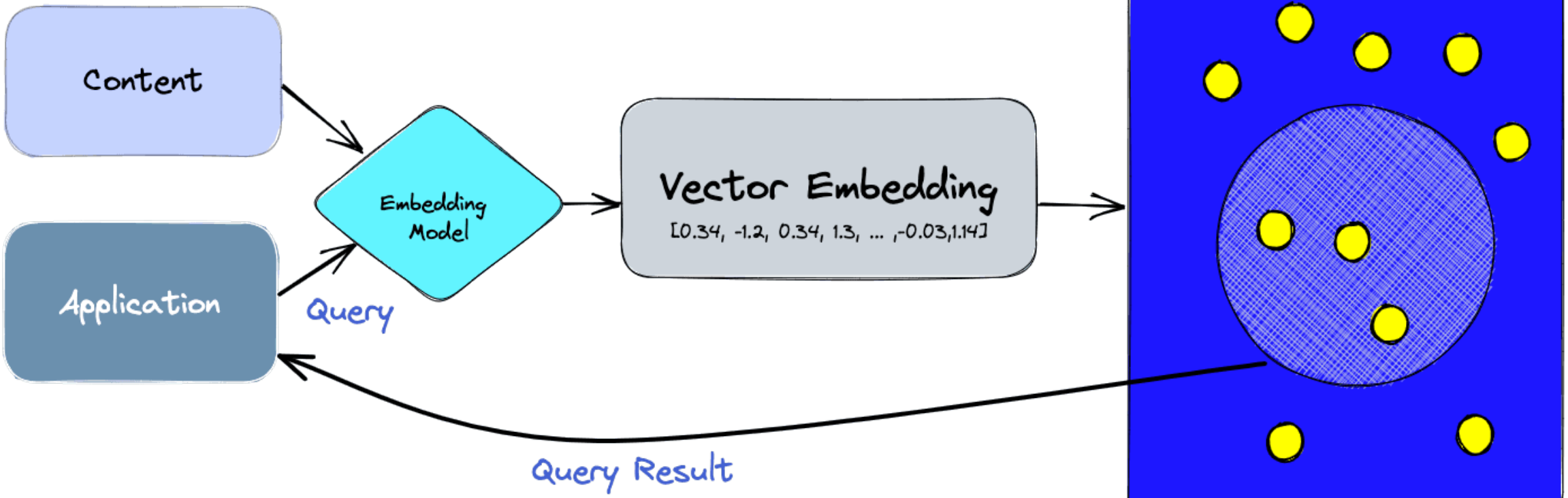
The Rise of Vector Databases



RSAC 2025 Conference

The Embedding Pipeline

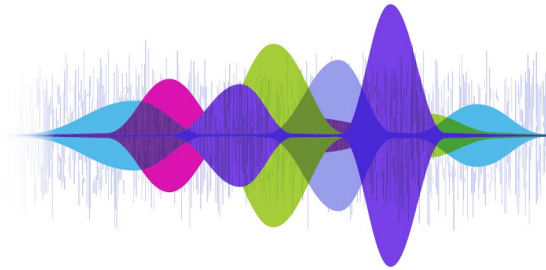
If these are compromised...



... all applications based on vector DBs are compromised

RSAC 2025 Conference

Word Embeddings



GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
jpennin@stanford.edu, richard@socher.org, manning@stanford.edu

GloVe

Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities using vector arithmetic, but the origin of these regularities has remained opaque. We analyze and make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global log-bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition.

42K citations

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

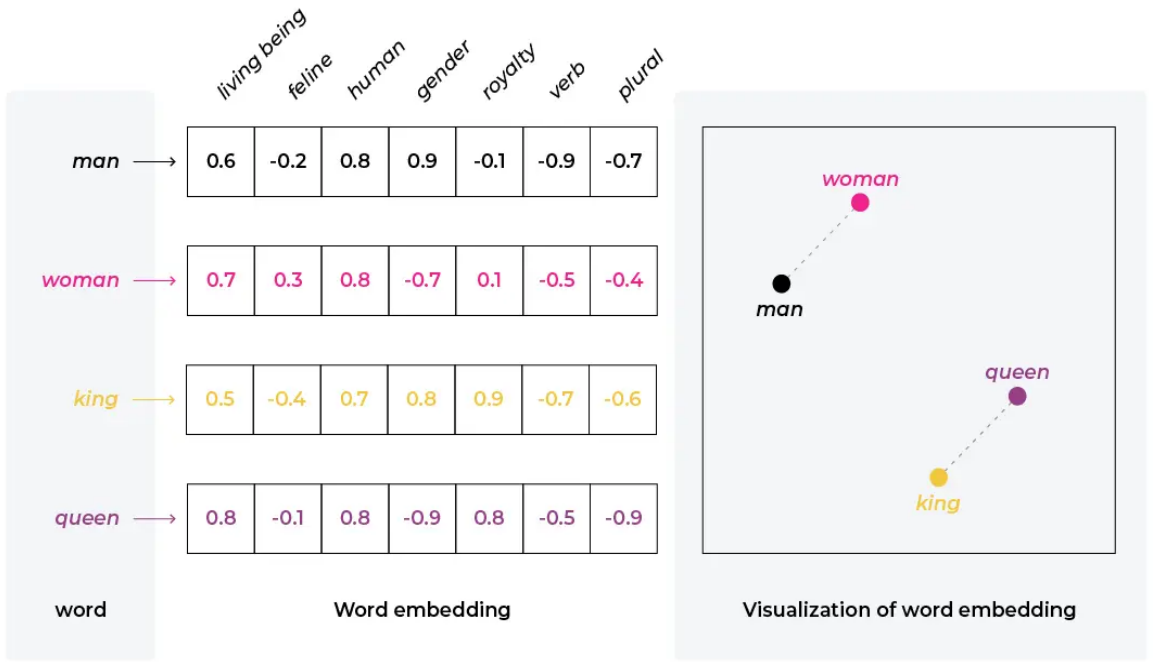
Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Word2vec

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.



Words → Vector representations

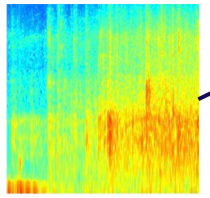
RSAC 2025 Conference

Key Concept: Alignment

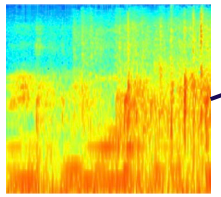
“Dog”

...

“Cat”



“Barking”



“Meowing”

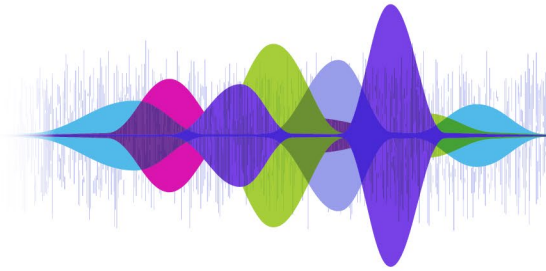
Multi-modal encoders

ImageBind

Titan

Vertex

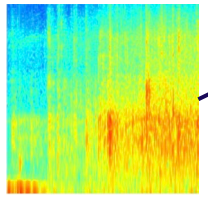
Key Concept: Alignment



Note: not to be confused with "safety alignment"

"Dog"

...

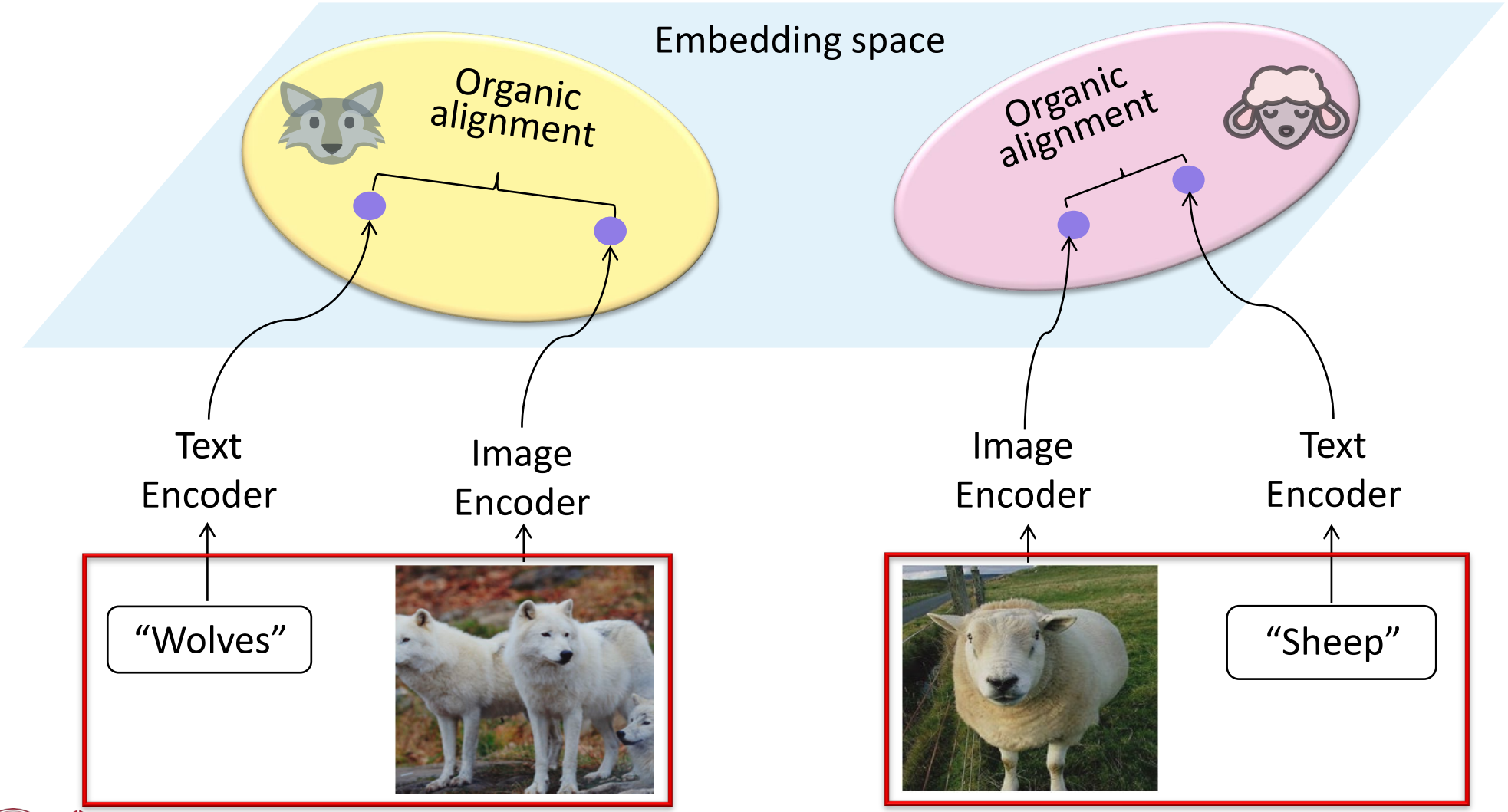


"Barking"

Semantically similar inputs are encoded into similar vectors

RSAC 2025 Conference

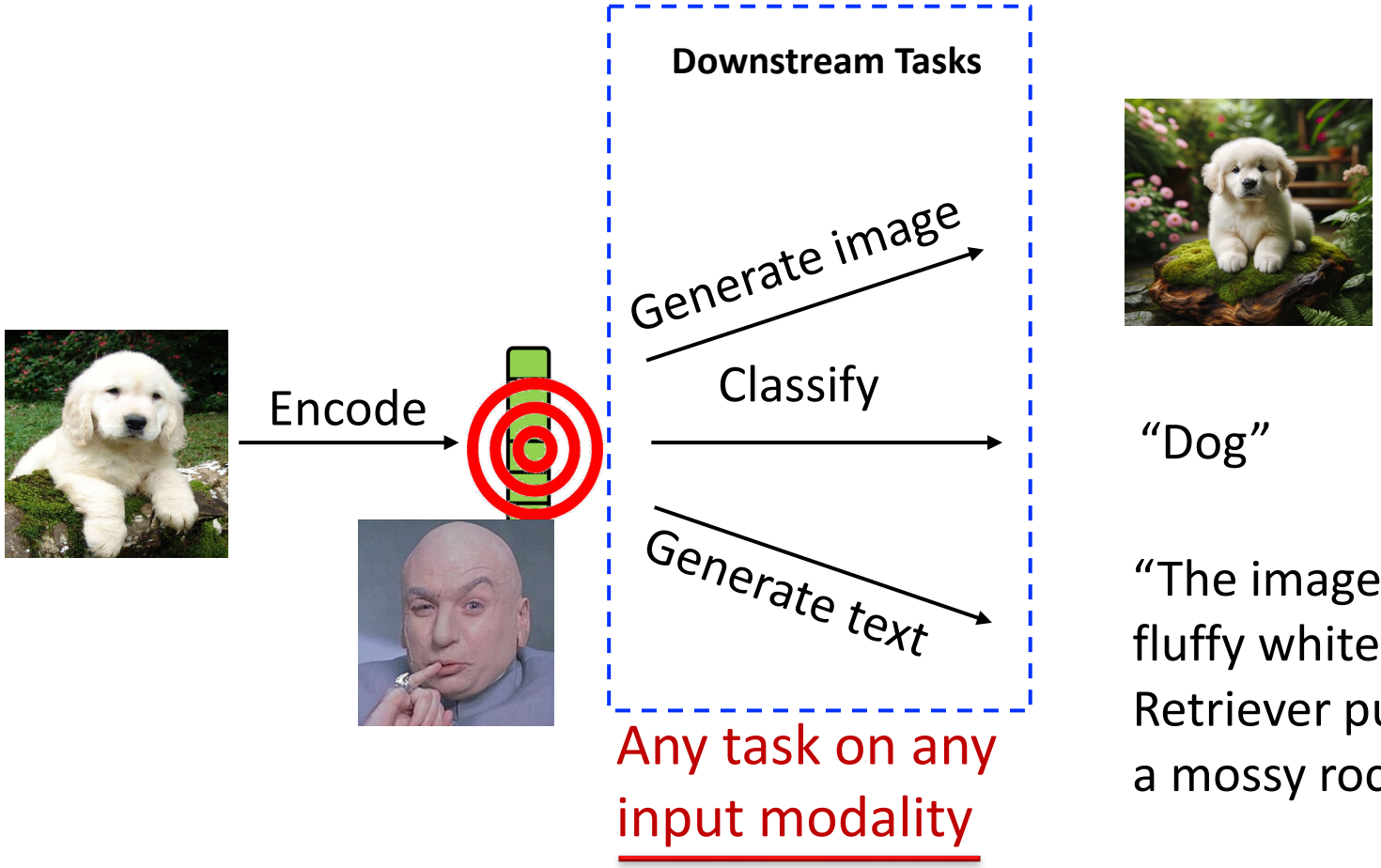
Alignment



semantically related inputs

semantically related inputs

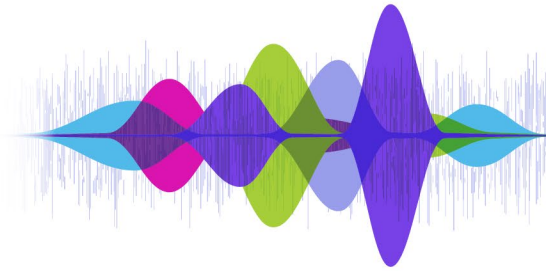
Embedding Alignment = Meaning



Even modalities the task was NOT trained on

RSAC 2025 Conference

Threat Model: Indirect Injection

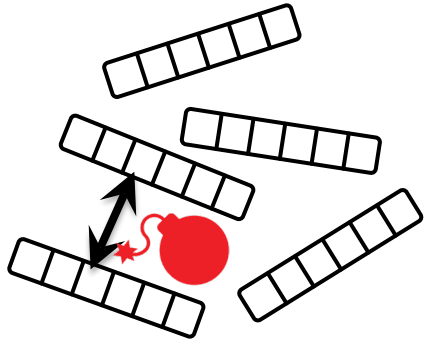


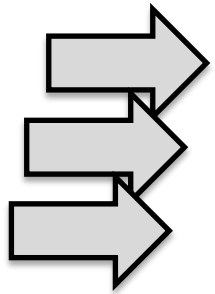
Text, images, sounds…

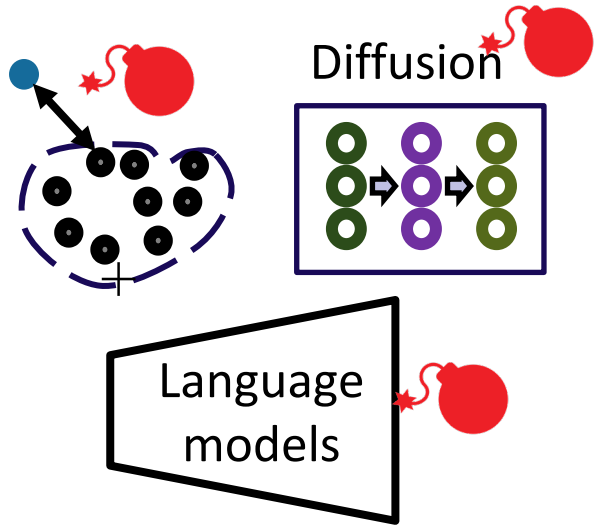
Vector embeddings

Retrieval, classification, generation

Encoders





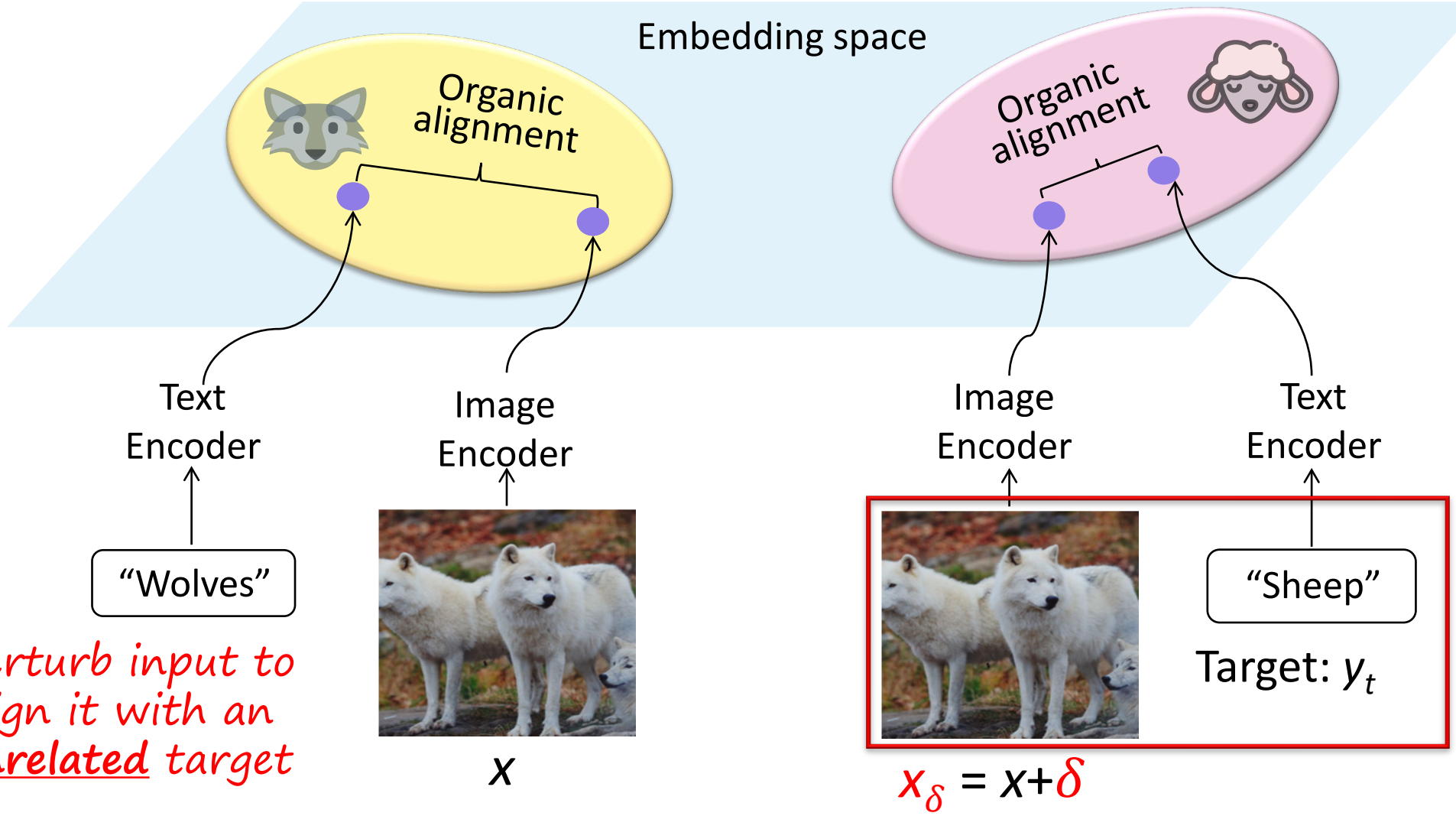


website, social media,
any user-generated content…

Embeddings are the only source of input semantics

RSAC 2025 Conference

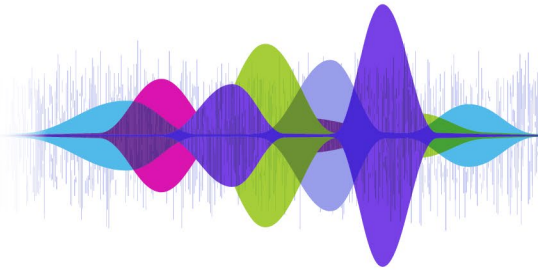
Adversarial Alignment



Perturb input to align it with an **unrelated** target

RSAC 2025 Conference

Multi-Modal Adversarial Illusions



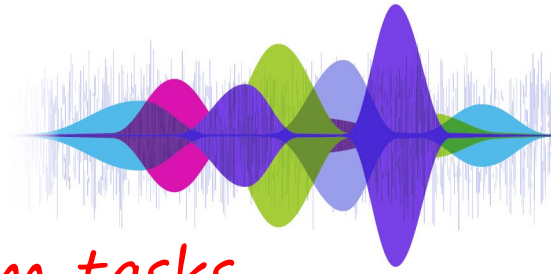
Clean input Alignment Target

$$\theta^{\mathrm{m}}(x+\delta) \sim \theta^{\overline{m}}(y_t)$$

Encoder for the input modality perturbation Encoder for the target modality

RSAC 2025 Conference

Turning Wolves into Sheep



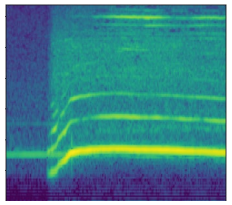
Works for all downstream tasks

Cross-modal

Downstream Tasks		
Generate image	Zero-shot (on ImageNet)	Can you describe this sound?

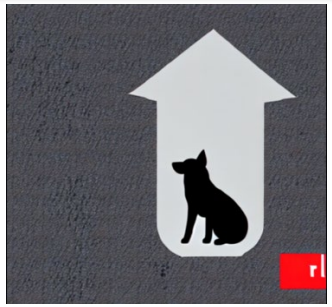
Image

Audio



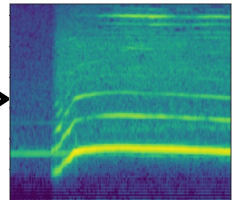
Wolf howling

Align in embedding space



White wolf, arctic wolf

"*The sound is a dog barking.*"



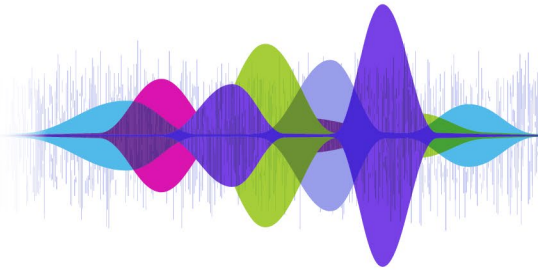
Ram, tup (sheep)

"*The sound of a sheep bleating.*"

Text (not one of the input modalities)

RSAC 2025 Conference

Multi-Modal Adversarial Illusions



Aren't these just adversarial examples?

- Different target — embedding alignment!
- Task agnostic
- Cross-modal

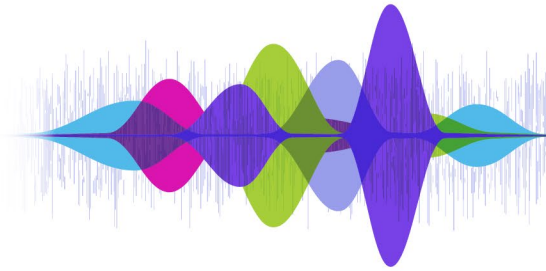
For example, use text to attack image-only models

- Adversarial alignment >>> organic alignment

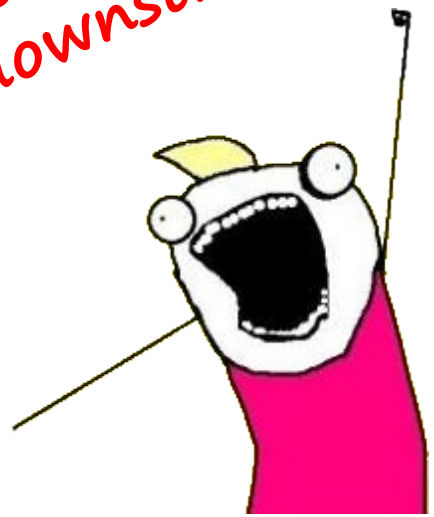
Defenses??

RSAC 2025 Conference

Multi-Modal Adversarial Illusions

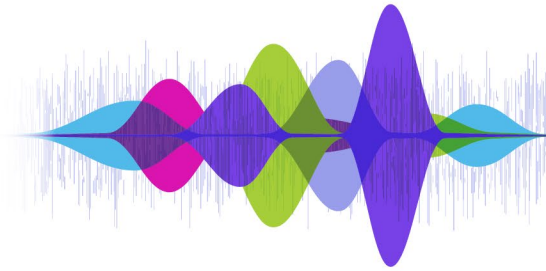


Compromise ALL downstream tasks



Align any input with any target

Schadenfreude



Downstream Tasks

Can you describe this image?

Generate image

"A man in prison cell"

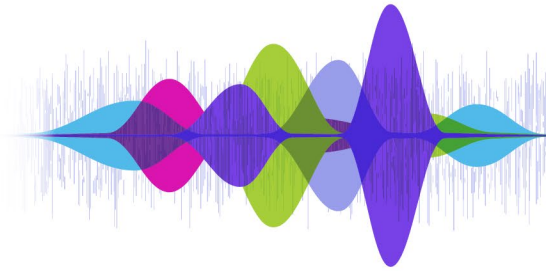
"The image features a man wearing a suit and tie, standing in a group of people. He is the only one wearing a suit."

Align in embedding space

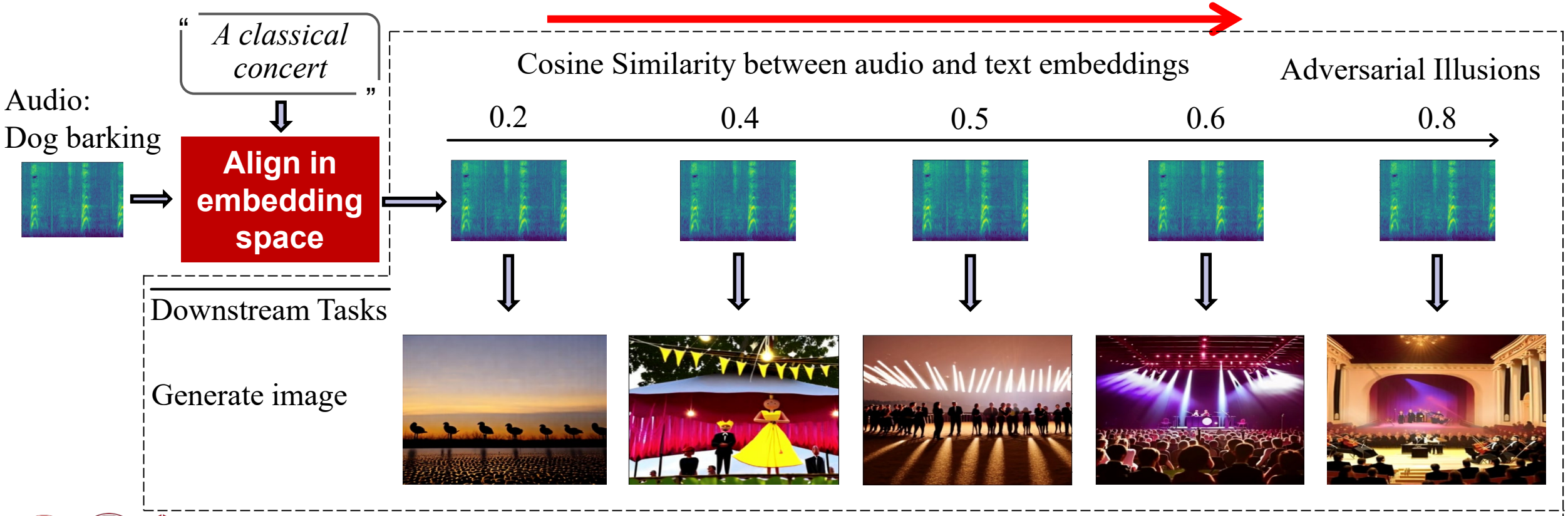
"The image shows a man standing behind a fence, peeking through a hole in the gate. He is wearing a white shirt and black pants."

RSAC 2025 Conference

A Symphony of Woofs

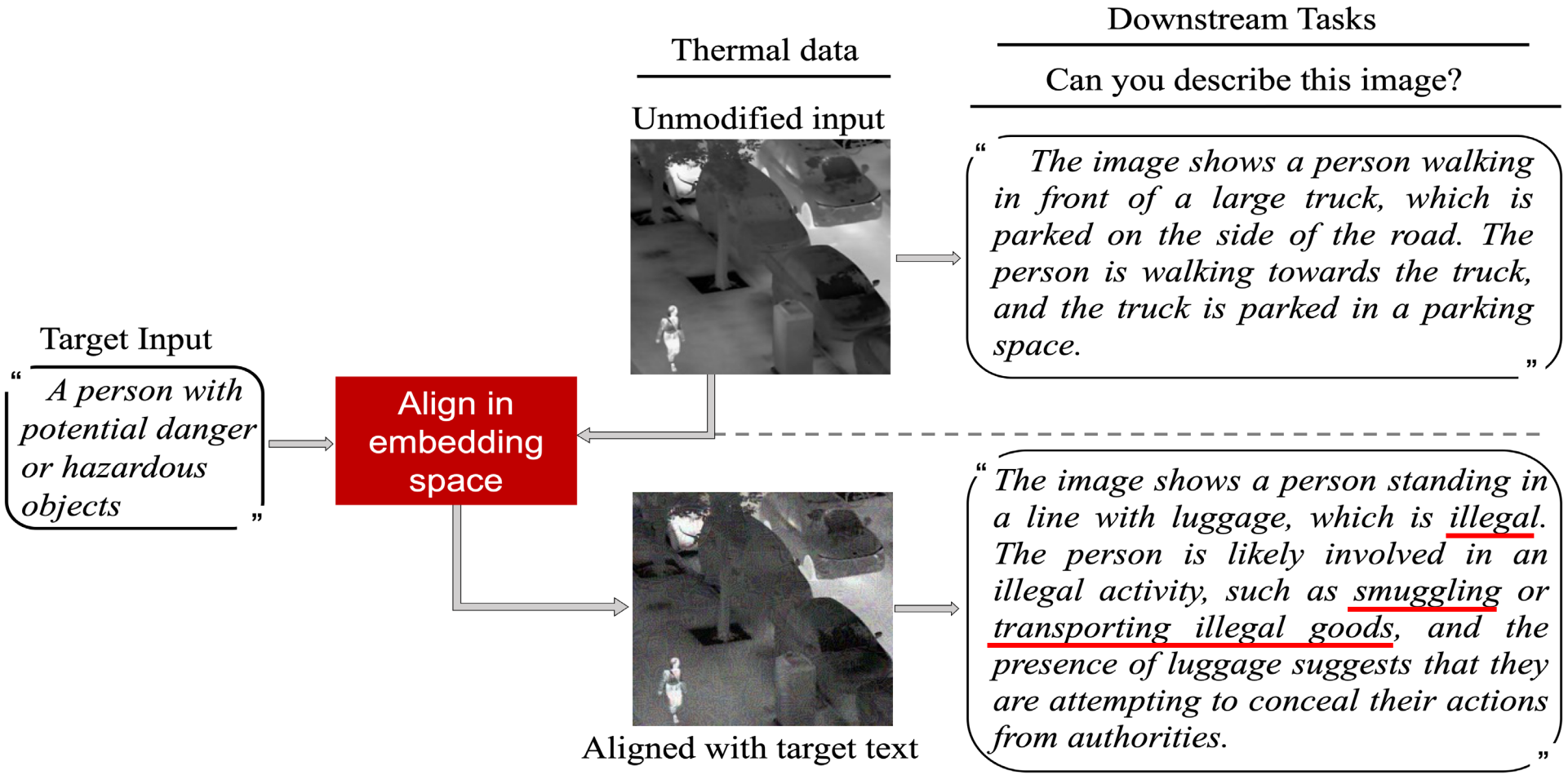


As alignment increases, the "meaning" of the input get closer to the adversary's target.



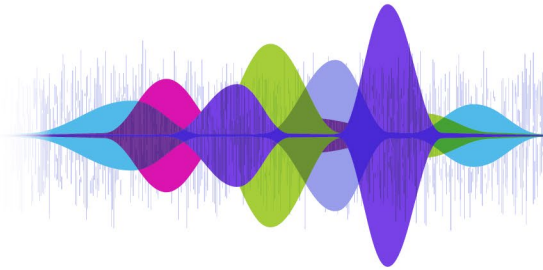
RSAC 2025 Conference

Surveillance


Thermal data
Unmodified input
Downstream Tasks
Can you describe this image?
"The image shows a person walking in front of a large truck, which is parked on the side of the road. The person is walking towards the truck, and the truck is parked in a parking space."
Target Input
"A person with potential danger or hazardous objects"
Align in embedding space
"The image shows a person standing in a line with luggage, which is illegal. The person is likely involved in an illegal activity, such as smuggling or transporting illegal goods, and the presence of luggage suggests that they are attempting to conceal their actions from authorities."
Aligned with target text

RSAC 2025 Conference

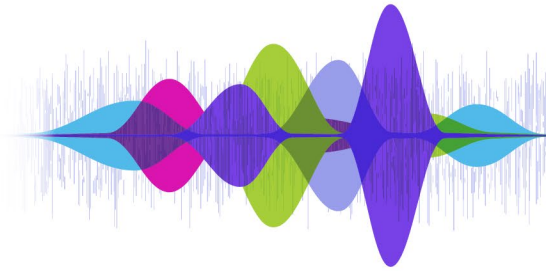
Adversary's Capabilities



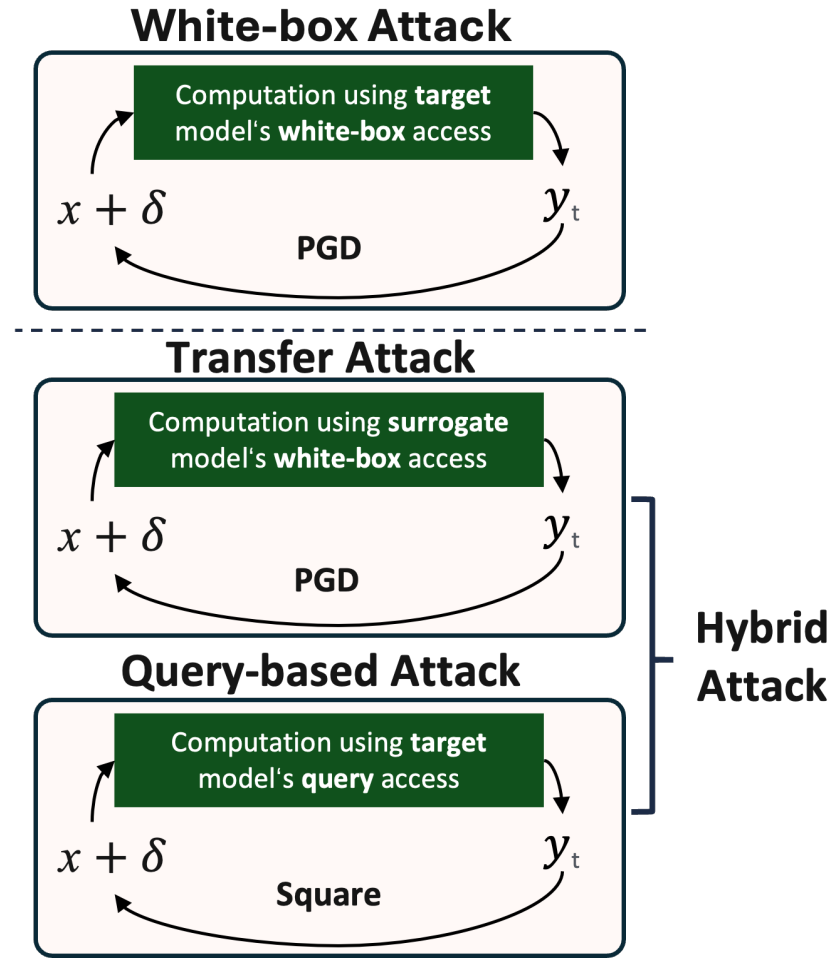
- White-box (full access to the target model)
- Black-box
  - Transfer (access to surrogate models only)
  - Query-based (can query the target model with limited queries)
  - Hybrid (Transfer + Query-based)

RSAC 2025 Conference

Craft Cross-Modal Illusions

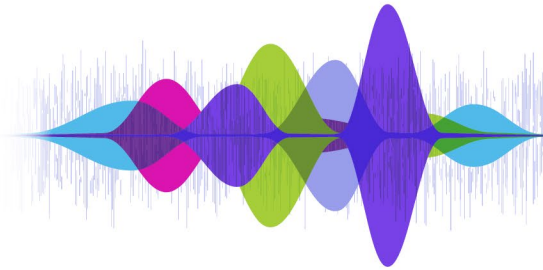


- **White-box**: iteratively update perturbation δ with PGD
- **Transfer**: generate illusions with local surrogate model
- **Query-based**: iteratively update perturbation δ with a variation of Square
- **Hybrid**: "warm-start" a query-based attack with locally generated illusion



RSAC 2025 Conference

White-Box Results



- **99%** success against zero-shot classification (images, thermal images, audio) and audio retrieval
- **68%** success against classification of generated text
- **64%** Top-1 success and **92%** Top-5 success against classification of generated images

If downstream models were better, attack would be more successful

RSAC 2025 Conference

Transfer Results

AudioCLIP:
adversarial alignment=0.2857

Transfer

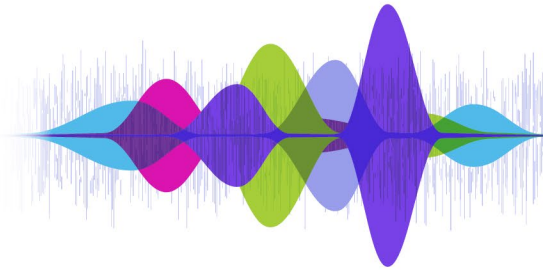
“ Centipede ”

ImageBind:
adversarial alignment=0.6784

Our illusions successfully fool all victim models with **97.5%** success rate.

RSAC 2025 Conference

Black-box Results



- **98%** success rates against black-box ImageBind and AudioCLIP with 18,942 and 4,112 queries (on average)
- **38%** Top-1 success and **58%** Top-5 success against classification of generated images with 100,000 queries

Amazon's Titan Embedding

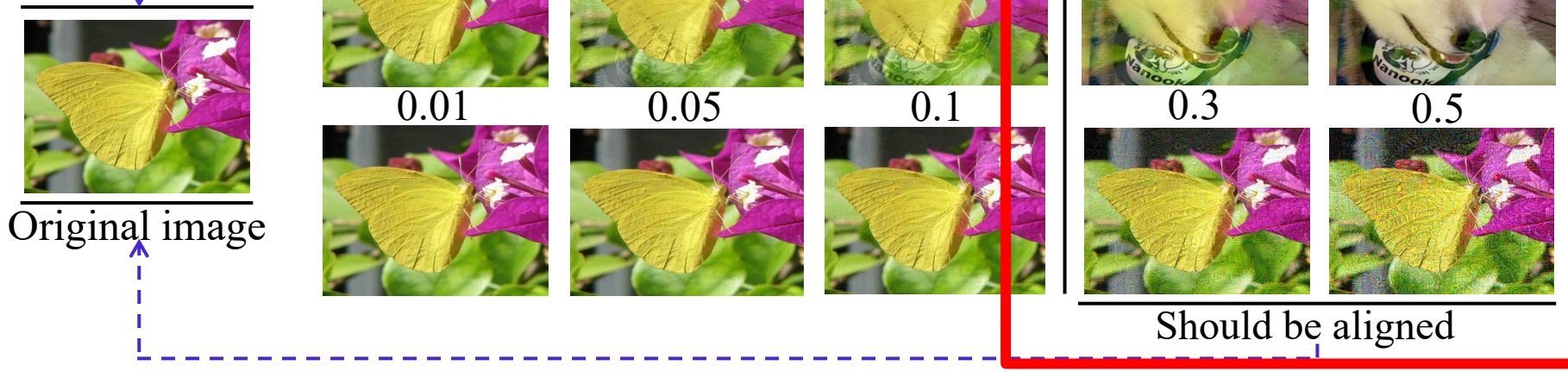
Commercial, proprietary, black-box embedding

- 30% success against zero-shot classification with 20,919 queries
- Hybrid attack: 42% success with 18,019 queries

RSAC 2025 Conference

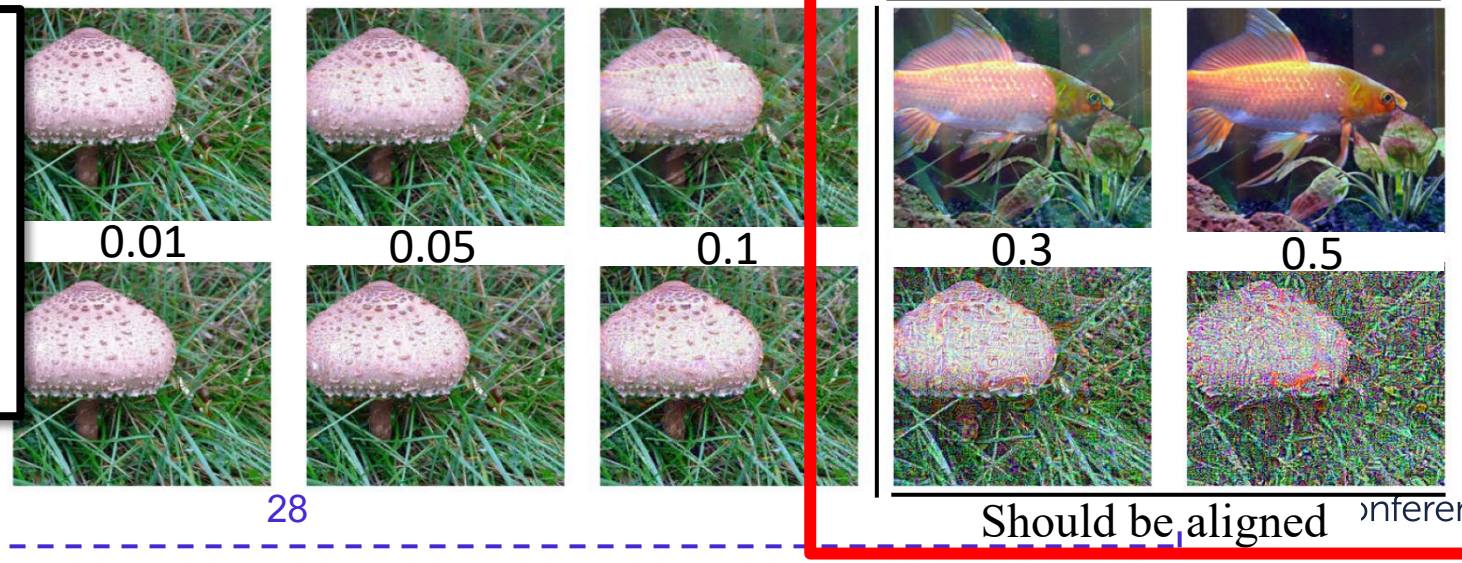
Certified Robustness

Force alignment between all inputs within small distance

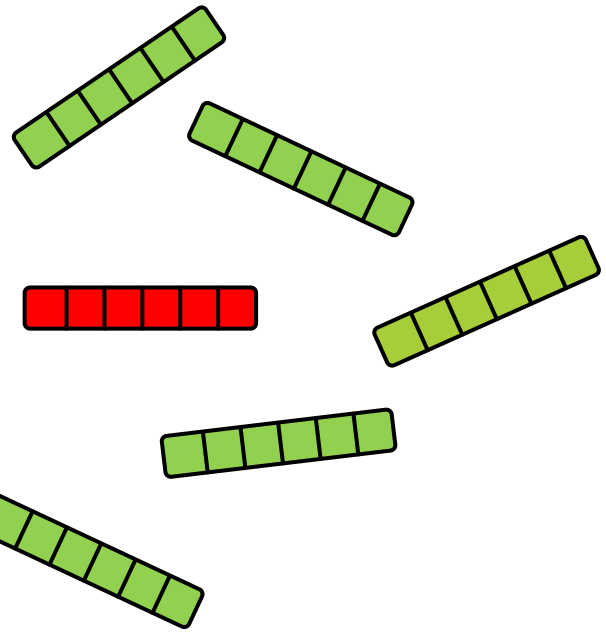


These images have the same distance from the original...

should "robust" embedding align them or not??

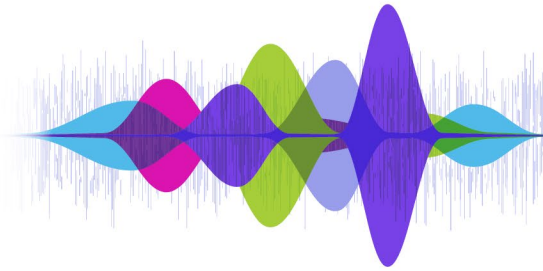


Adversarial Alignment on Steroids



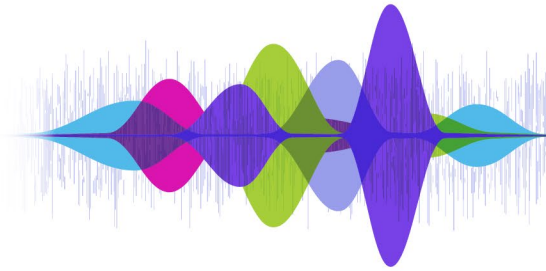
RSAC 2025 Conference

Why?



- Search engine optimization
- Spamming retrieval and search
- Spamming product listings
- Spreading misinformation

Motivations



OII > NEWS & EVENTS > NEWS >

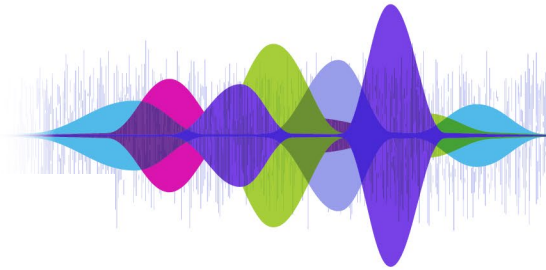
PRESS RELEASE -

Junk news sites using advertising and web search optimization to spread COVID-19 misinformation, finds new study

Published on
3 Aug 2020

A new report from the Oxford Internet Institute (OII), in collaboration with Oxford Information Labs (OxIL) has found that more than 60% of junk news sites are using one of the major advertising platforms help convert website traffic into income.

RSAC 2025 Conference



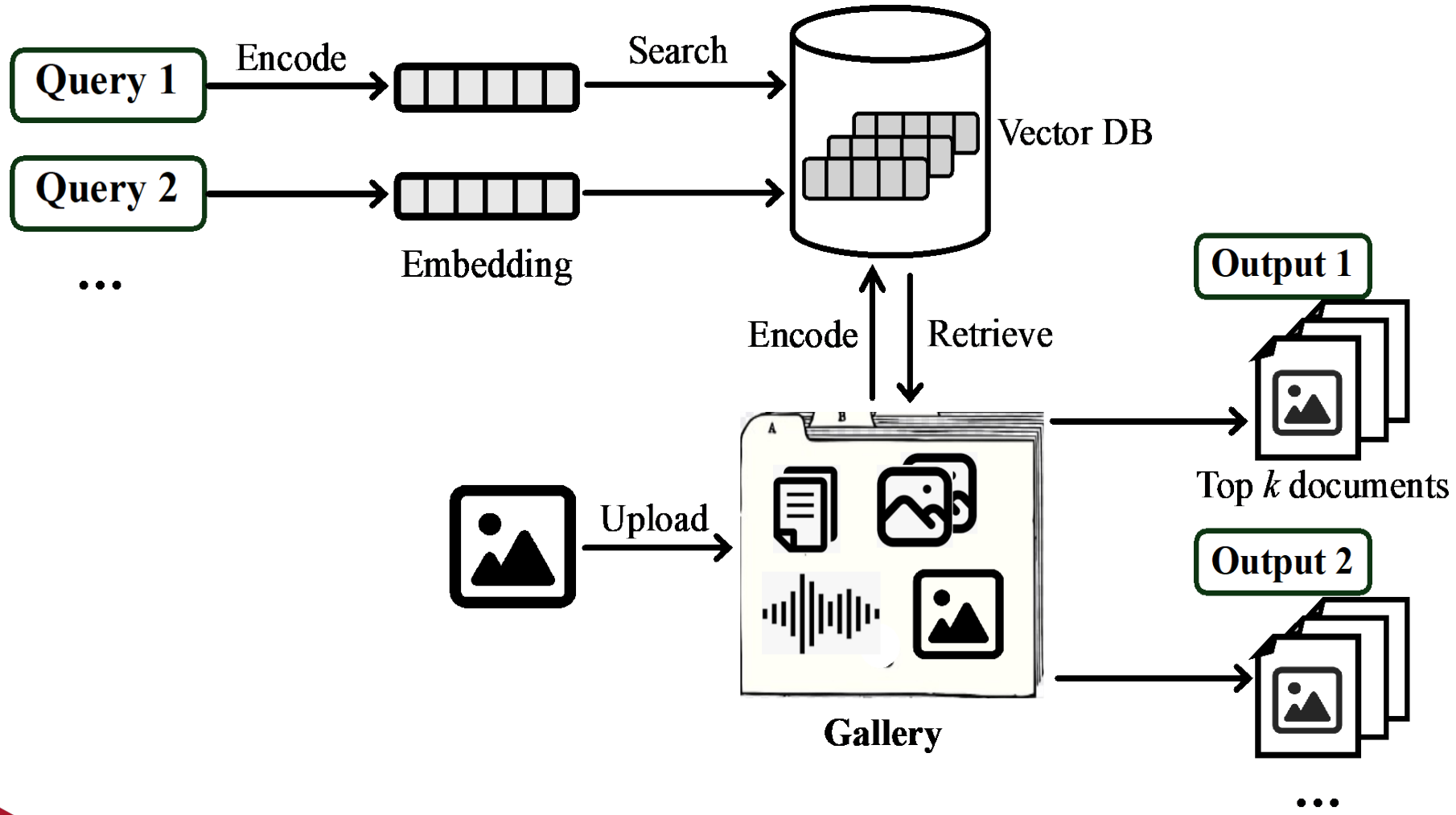
“People upload massive amounts of albums that are intended to be streaming fraud albums.”

Another strategy is creating AI covers of popular songs and getting them onto popular playlists so normal people will stream them. Another involves bots “listening” to songs.

Earlier this year, a Danish man was sentenced to 18 months in prison for using bots to get about $300,000 in royalties. Another man, Michael Smith, was arrested and charged with defrauding streaming services out of $10 million over the course of seven years. Smith used AI tools to create hundreds of thousands of songs under the names of fake artists such as “Calm Baseball,” “Calm Connected,” and “Calm Knuckles.” He then streamed the huge catalog using bots, billions of times, prosecutors allege. That diverted money that should have gone to real musicians that real people were really listening to. (In this case, Spotify paid only $60,000 to Smith, suggesting the company’s protective measures worked to limit payments, Macowski said.)

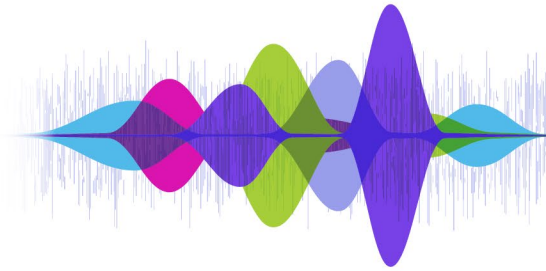
RSAC 2025 Conference

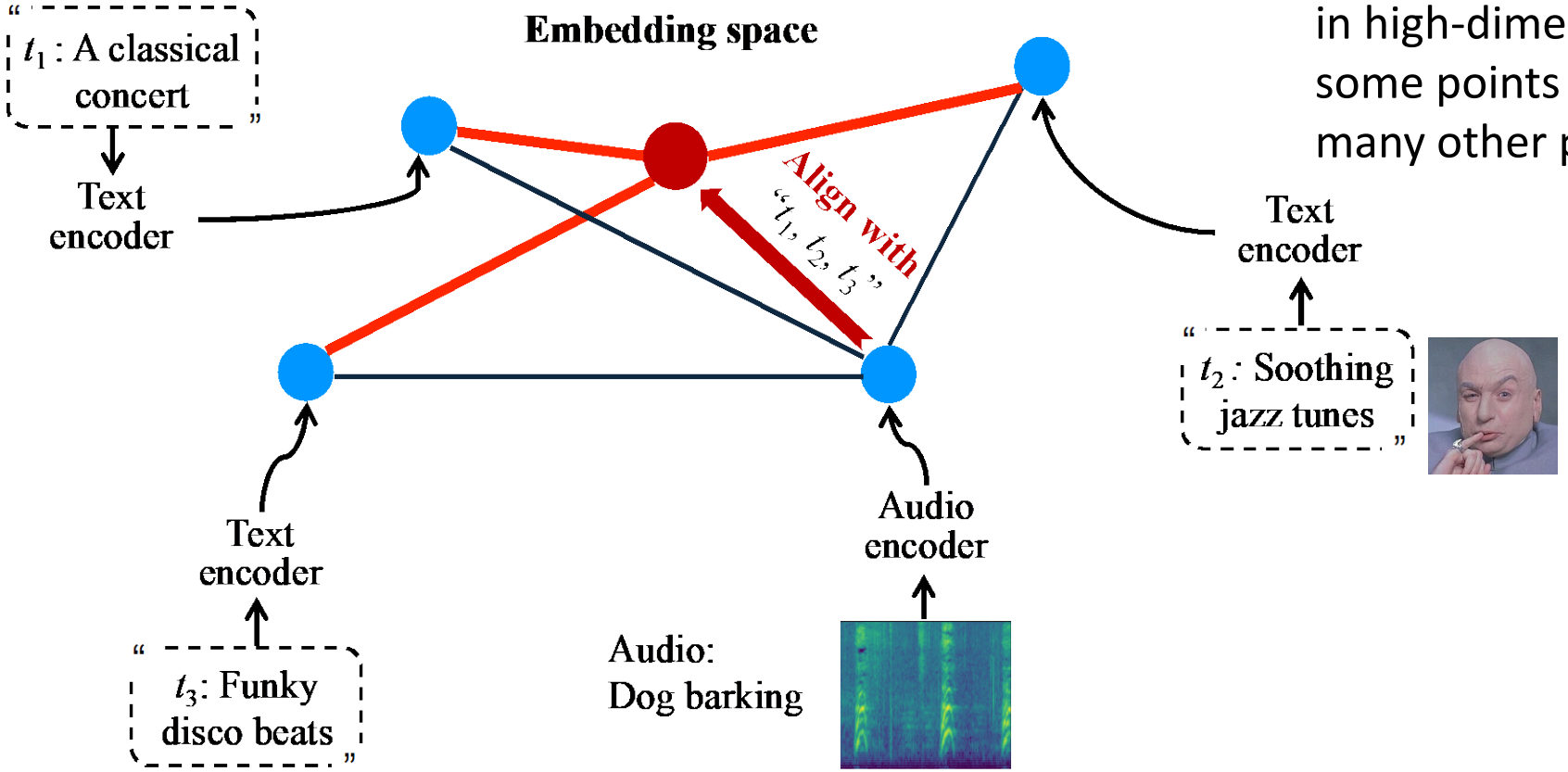
Cross-Modal Retrieval



RSAC 2025 Conference

Exploiting Hubness



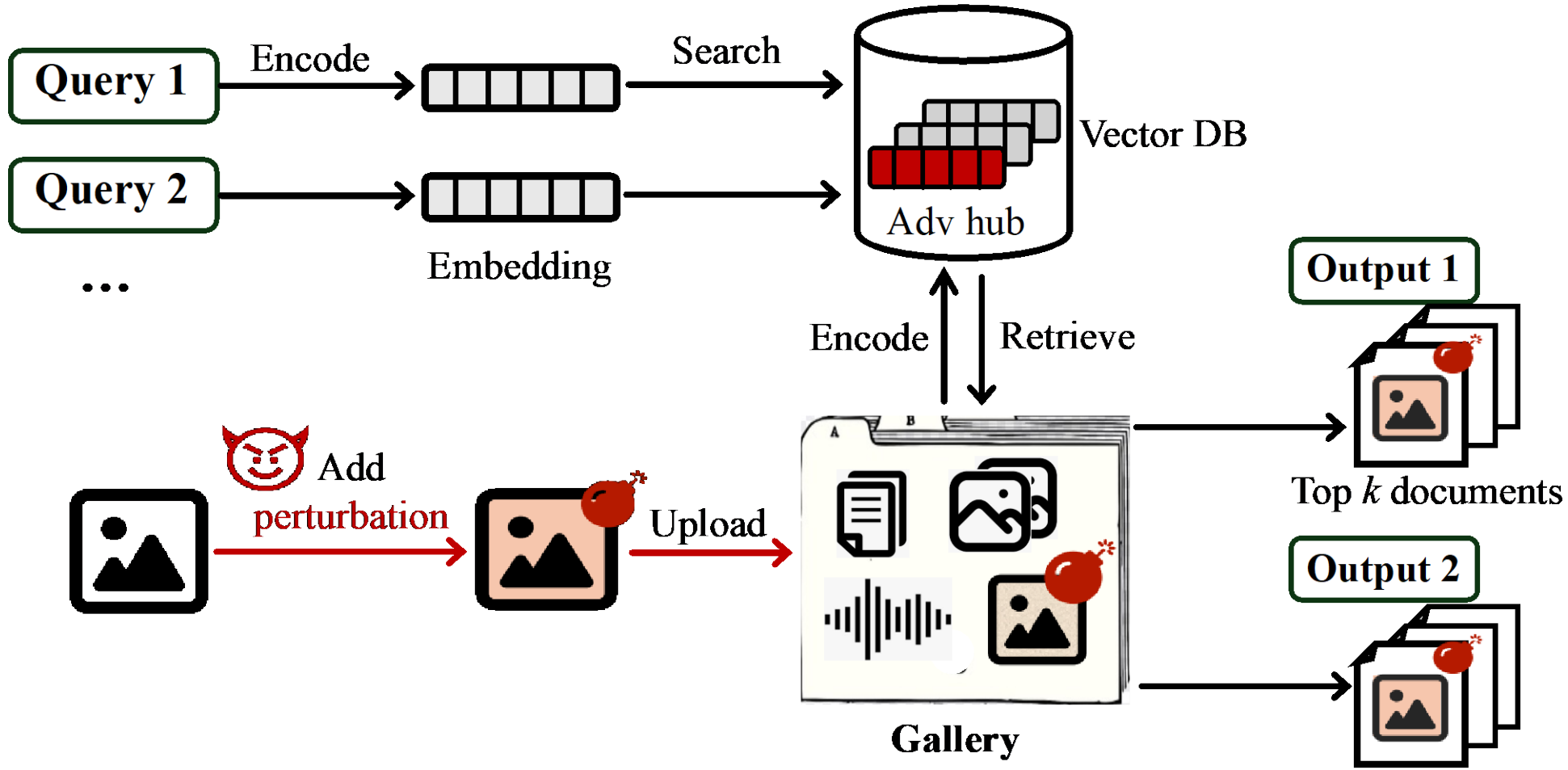


Well-known phenomenon in information retrieval:
in high-dimensional vector spaces, some points appear as neighbors of many other points

Perturb input to align with many targets at once, i.e., create an artificial "hub"

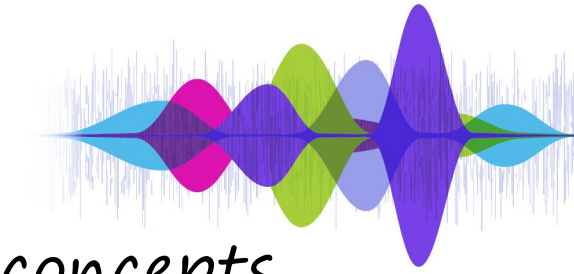
RSAC 2025 Conference

Attacking Cross-Modal Retrieval



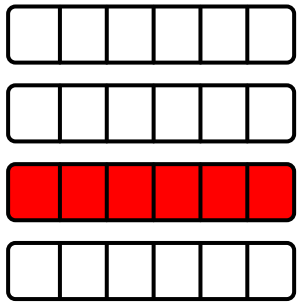
RSAC 2025 Conference

Universal Spam



Can also target specific concepts

Looks like a rele



"People upload massive amounts of albums that are intended to be streaming fraud albums."

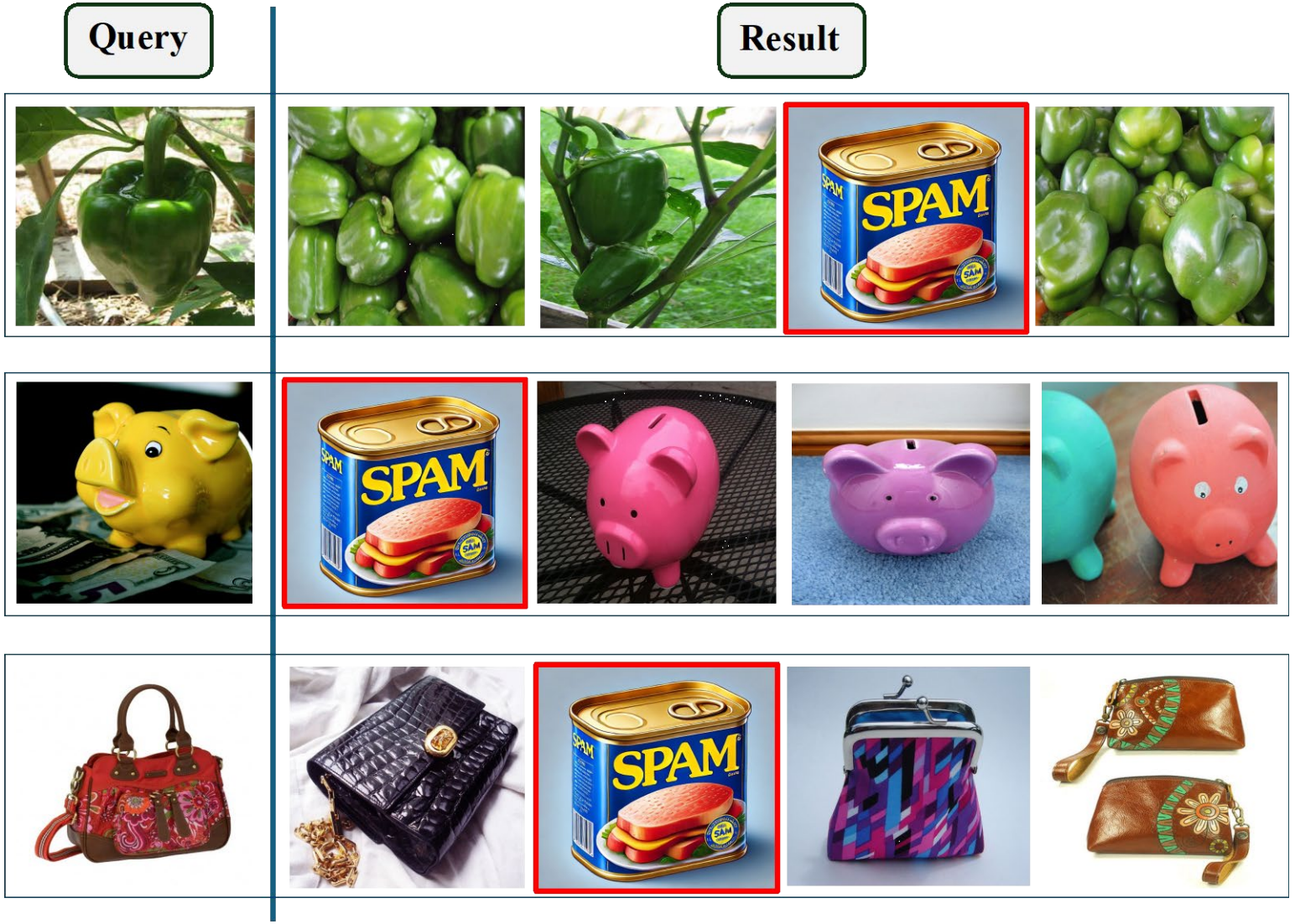
Another strategy is creating AI covers of popular songs and getting them onto popular playlists so normal people will stream them. Another involves bots "listening" to songs.

Earlier this year, a Danish man was sentenced to 18 months in prison for using bots to get about $300,000 in royalties. Another man, Michael Smith, was arrested and charged with defrauding streaming services out of $10 million over the course of seven years. Smith used AI tools to create hundreds of thousands of songs under the names of fake artists such as "Calm Baseball," "Calm Connected," and "Calm Knuckles." He then streamed the huge catalog using bots, billions of times, prosecutors allege. That diverted money that should have gone to real musicians that real people were really listening to. (In this case, Spotify paid only $60,000 to Smith, suggesting the company's protective measures worked to limit payments, Macowski said.)

These threats are real!

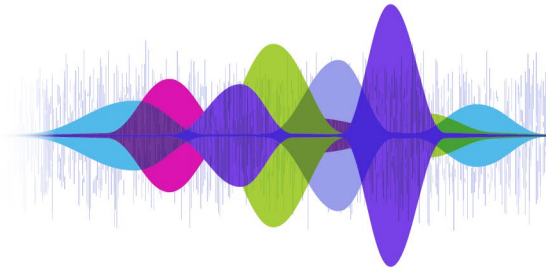
RSAC 2025 Conference

Image-to-image retrieval

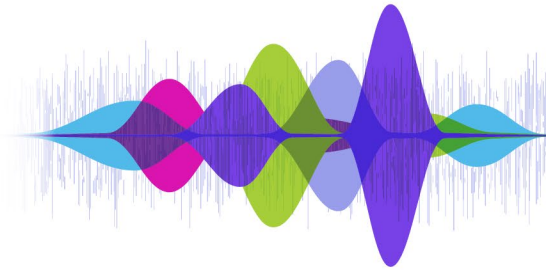


RSAC 2025 Conference

Text-to-image retrieval



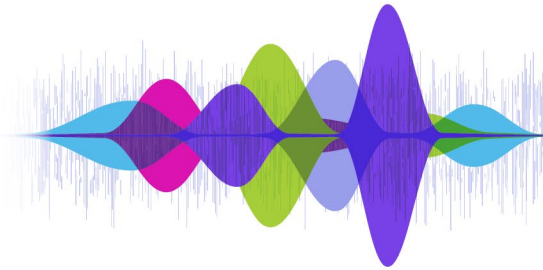
Results



- 82.3% top-1 success against text-to-image retrieval, aligning one image with 25k captions
- 56.9% top-5 success against image-to-image retrieval, 31.7% success against Pinecone image search (black-box)
- 83.6% top-5 success against text-to-audio retrieval

Only used 100 randomly selected query sample!

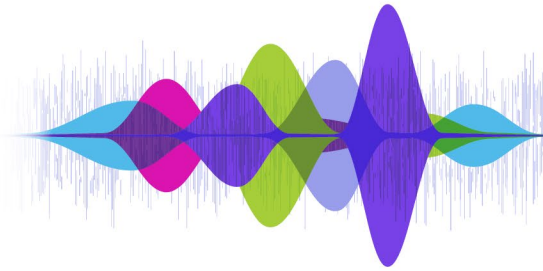
Takeaways



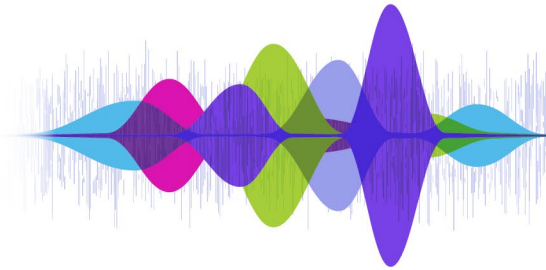
- Multi-modal embeddings and LLMs are juicy targets for cross-modal, task-agnostic attacks
  - Retrieval (adversarial hubness), zero-shot classification, generation
  - Adversary need not know the task or even which modalities the task accepts (text, images, audio, thermal images…)

RSAC 2025 Conference

Apply What You Have Learned



- Multi-modal AI today is not ready for tasks where correctness is important:
  - Errors can lead to serious consequences in fields like medicine, law, or safety-critical systems.
- If you must use multi-modal AI, test and red-team it:
  - Rigorously test AI on realistic, adversarial, and edge-case inputs.



What have we learned from 10 years of research and 10 million papers on adversarial robustness?

RSAC™ | 2025 Conference

Many Voices.
One Community.

Thank You

Our code is available!